CONTENTS

5 Generalized Linear Models

5.1 Introduction

The two main tools of statistical inference are confidence intervals and hypothesis tests.

Confidence intervals, also known as interval estimates, are increasingly regarded as more useful than hypothesis tests because the width of a confidence interval provides a measure of the precision with which inferences can be made. It does so in a way which is conceptually simpler than the power of a statistical test.

Hypothesis tests in a statistical modelling framework are performed by comparing how well two related models fit the data.

If the response variables are Normally distributed, the sampling distributions used for inference can often be determined exactly. For other distri-

butions we need to rely on large-sample asymptotic results based on the Central Limit Theorem. The rigorous development of these results requires careful attention to various regularity conditions. For independent observations from distributions which belong to the exponential family and in particular for generalized linear models, the necessary conditions are indeed satisfied.

The basic idea is that under appropriate conditions, if S is a statistic of interest, then approximately

$$\frac{(S - E(S))}{\sqrt{(Var(S))}} \sim N(0, 1)$$

or equivalently

$$\frac{(S - E(S))^2}{(Var(S))} \sim \chi^2(1)$$

where $E(S)$ and $Var(S)$ are the expectation and variance of S, respectively.

If there is a vector of statistics of interest $\mathbf{s} =$

$\begin{bmatrix} S_1 \\ \vdots \\ S_p \end{bmatrix}$ with asymptotic expectation $E(\mathbf{S})$ and asymptotic variance covariance matrix $\mathbf{V}$, then approximately

$$[\mathbf{S}-E(\mathbf{S})]^T\mathbf{V}^{-1}[\mathbf{S}-E(\mathbf{S})] \sim \chi^2(p) \tag{5.1}$$

provided $\mathbf{V}$ is non-singular so a unique inverse matrix $\mathbf{V}^{-1}$ exists.

5.2 Sampling distribution for score statistics

Suppose $Y_1, \ldots, Y_N$ are independent random variables in a generalized linear model with parameters $\boldsymbol{\beta}$, where $E(Y_i) = \mu_i$ and $g(\mu_i) = x_i^T\boldsymbol{\beta} = \eta_i$.

The score statistics are

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N}\left[\frac{(Y_i - \mu_i)}{Var(Y_i)}x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]$$

$$\text{for } j = 1, \ldots, p.$$

As $E(Y_i) = \mu_i$ for all i, $E(U_j) = 0$ for $j = 1, \ldots, p$.

The variance covariance matrix of the score statistics is the information matrix $\mathcal{J}$ with elements

$$\mathcal{J}_{jk} = E[U_j U_k] = -E\left[\frac{\partial l^2}{\partial \beta_j \partial \beta_k}\right] = -E\left[\frac{\partial U_j}{\partial \beta_k}\right].$$

If there is only one parameter β, the score statistic has the asymptotic sampling distribution

$$\frac{U}{\sqrt{\mathcal{J}}} \sim N(0,1) \text{ or } \frac{U^2}{\mathcal{J}} \sim \chi^2(1).$$

If there is a vector of parameters

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

then the score vector

$$\mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix}$$

has the multivariate Normal distribution $\mathbf{U} \sim MVN(\mathbf{0}, \mathcal{J})$, at least asymptotically, and so

$$\mathbf{U}^T \mathcal{J}^{-1} \mathbf{U} \sim \chi^2(p)$$

for large samples.

Example 1.
If $Y \sim Bin(n, \pi)$.

(i) Find the score statistic, U.

(ii) Find the variance of U, $\mathcal{J}$.

(iii) Find the asymptotically distribution of $\frac{U}{\sqrt{\mathcal{J}}}$.

Example 2.
You are given the following probability density function for a single random variable, X:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp(-z^2/2), z = \frac{\ln x - \mu}{\sigma}.$$

(a) Find the score function, $\mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$.

(b) Find the information matrix, $\mathcal{J}$.

(c) Find the asymptotic distribution of $\mathbf{U}$.

5.3 Taylor series approximations

To obtain the asymptotic sampling distributions for various other statistics, it is useful to use Taylor series approximations. The Taylor series approximation for a function $f(x)$ of a single variable x about a value t is

$$f(x) = f(t)+(x-t)\left[\frac{df}{dx}\right]_{x=t}+\frac{1}{2}(x-t)^2\left[\frac{d^2f}{dx^2}\right]_{x=t}+\cdots$$

provided that x is near t.

For a log-likelihood function of a single parameter β , the first three terms of the Taylor series approximation near an estimate β are

$$l(\beta) = l(b) + (\beta - b)U(b) + \frac{1}{2}(\beta - b)^2U'(b)$$

where $U(b) = \frac{dl}{d\beta}$ is the score function evaluated at $\beta = b$. If $U'(b) = \frac{d^2l}{d\beta^2}$ is approximated by its expected value $E(U') = -\mathcal{J}$, the approximation

becomes

$$l(\beta) = l(b) + (\beta - b)U(b) - \frac{1}{2}(\beta - b)^2\mathcal{J}(b)$$

where $\mathcal{J}(b)$ is the information evaluated at $\beta = b$.

The corresponding approximation for the log-likelihood function for a vector parameter $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}) = l(\mathbf{b})+(\boldsymbol{\beta}-\mathbf{b})^T\mathbf{U}(\mathbf{b})^{-1/2}(\boldsymbol{\beta}-\mathbf{b})^T\mathcal{J}(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b})$$

where $\mathbf{U}$ is the vector of scores and $\mathcal{J}$ is the information matrix.

For the score function of a single parameter β, the first two terms of the Taylor series approximation near an estimate b give

$$U(\beta) = U(b) + (\beta - b)U'(b).$$

If U' is approximated by $E(U') = -\mathcal{J}$ then

$$U(\beta) = U(b) - (\beta - b)\mathcal{J}(b).$$

The corresponding expression for a vector parameter $\boldsymbol{\beta}$ is

$$U(\boldsymbol{\beta}) = U(\mathbf{b}) - \mathcal{J}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b}). \qquad (5.2)$$

5.4 Sampling distribution for maximum likelihood estimators

Equation (5.2) can be used to obtain the sampling distribution of the maximum likelihood estimator $\widehat{\boldsymbol{\beta}} = \mathbf{b}$. By definition, $\mathbf{b}$ is the estimator which maximizes $l(\mathbf{b})$ and so $U(\mathbf{b}) = 0$. Therefore, $U(\boldsymbol{\beta}) = -\mathcal{J}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})$, or equivalently, $(\mathbf{b} - \boldsymbol{\beta}) = \mathcal{J}^{(-1)}\mathbf{U}$ provided that $\mathcal{J}$ is non-singular.

If $\mathcal{J}$ is regarded as constant, then $E(\mathbf{b} - \boldsymbol{\beta}) = 0$ because $E(\mathbf{U}) = 0$. Therefore, $E(\mathbf{b}) = \boldsymbol{\beta}$, at least asymptotically, so $\mathbf{b}$ is a consistent estimator of $\boldsymbol{\beta}$. The variance covariance matrix for $\mathbf{b}$ is

$$E[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})^T] = \mathcal{J}^{-1}E(\mathbf{U}\mathbf{U}^T)\mathcal{J}^{-1} = \mathcal{J}^{-1}.$$

Because $\mathcal{J} = E(\mathbf{U}\mathbf{U}^T)$ and $(\mathcal{J}^{-1})^T = \mathcal{J}^{-1}$ as $\mathcal{J}$ is symmetric.
The asymptotic sampling distribution for $\mathbf{b}$, is

$$(\mathbf{b} - \boldsymbol{\beta})^T\mathcal{J}(\mathbf{b} - \boldsymbol{\beta}) \sim \chi^2(p).$$

This is the Wald statistic.

For the one-parameter case, the more commonly used form is $b \sim N(\beta, \mathcal{J}^{-1})$.

5.5 Log-likelihood ratio statistic

One way of assessing the adequacy of a model is to compare it with a more general model with the maximum number of parameters that can be estimated.

- **Saturated model** - A more general model with the maximum number of parameters that can be estimated.
- **maximal** or **full** model - If there are N observations $Y_i, i = 1, \ldots, N$, all with potentially different values for the linear component $\mathbf{x}_i^T\boldsymbol{\beta}$, then a saturated model can be specified with N parameters.
- **Replicates** - observations that have the same linear component or covariate pattern, that is,

they correspond to the same combination of factor levels and have the same values of any continuous explanatory variables. In this case, the maximum number of parameters that can be estimated for the saturated model is equal to the number of potentially different linear components, which may be less than N.

In general, let

- m denote the maximum number of parameters that can be estimated.
- $\boldsymbol{\beta}_{\max}$ denote the parameter vector for the saturated model.
- $\mathbf{b}_{\max}$ denote the maximum likelihood estimator of $\boldsymbol{\beta}_{\max}$.
- $L(\mathbf{b}_{\max}; y)$ denote the likelihood function for the saturated model evaluated at $\mathbf{b}_{\max}$.
- $L(\mathbf{b}; y)$ denote the maximum value of the likelihood function for the model of interest.

Note that $L(\mathbf{b}_{\max}; y)$, will be larger than any other likelihood function for these observations,

with the same assumed distribution and link function, because it provides the most complete description of the data.

The likelihood ratio

$$\lambda = \frac{L(\mathbf{b}_{\max}; y)}{L(\mathbf{b}; y)}$$

provides a way of assessing the goodness of fit for the model. In practice, the logarithm of the likelihood ratio, which is the difference between the log-likelihood functions,

$$\ln \lambda = l(\mathbf{b}_{\max}; y) - l(\mathbf{b}; y)$$

is used. Large values of $\ln \lambda$ suggest that the model of interest is a poor description of the data relative to the saturated model. To determine the critical region for $\ln \lambda$, its sampling distribution is needed.

5.6 Sampling distribution for the deviance

The deviance, also called the log-likelihood (ratio) stati is

$$2[l(\mathbf{b}_{\max}; y) - l(\mathbf{b}; y)]$$

If $\mathbf{b}$ is the maximum likelihood estimator of the parameter $\boldsymbol{\beta}$ (so that $U(\mathbf{b}) = 0$),

$$l(\boldsymbol{\beta}) - l(\mathbf{b}) = -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})^T \mathcal{J}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})$$

approximately. Therefore, the statistic

$$2[l(\mathbf{b}; y) - l(\boldsymbol{\beta}; y)] = (\boldsymbol{\beta} - \mathbf{b})^T \mathcal{J}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})$$

which has the chi-squared distribution $\chi^2(p)$, where p is the number of parameters.

From this result the sampling distribution for the deviance can be derived

$$\begin{aligned} 2[l(\mathbf{b}_{\max}; y) - l(\mathbf{b}; y)] = {} & 2[l(\mathbf{b}_{\max}; y) - l(\boldsymbol{\beta}_{\max}; y)] \\ & -2[l(\mathbf{b}; y) - l(\boldsymbol{\beta}; y)] \\ & +2[l(\boldsymbol{\beta}_{\max}; y) - l(\boldsymbol{\beta}; y)]. \end{aligned}$$

Note that,

- $2[l(\mathbf{b}_{\max}; y) - l(\boldsymbol{\beta}_{\max}; y)] \sim \chi^2(m)$
- $2[l(\mathbf{b}; y) - l(\boldsymbol{\beta}; y) \sim \chi^2(p)$, where p is the number of parameters in the model of interest.
- $2[l(\boldsymbol{\beta}_{\max}; y) - l(\boldsymbol{\beta}; y)]$ is a positive constant which will be near zero if the model of interest fits the data almost as well as the saturated model fits.

Therefore, the sampling distribution of the deviance is approximately

$$D \sim \chi^2(m - p, v)$$

where v is the non-centrality parameter. The deviance forms the basis for most hypothesis testing for generalized linear models.

t A good rule of thumb is to divide the deviance by its number of degrees of freedom. If the ratio $\frac{D}{m-p}$ is much greater than unity, the current model is not an adequate fit to the data.

If the response variables Y_i are Normally distributed, then D has a chi squared distribution exactly. In this case, however, D depends on $Var(Y_i) = \sigma^2$, which, in practice, is usually unknown. This means that D cannot be used directly as a goodness of fit statistic.

For $Y_i's$ with other distributions, the sampling distribution of D may be only approximately chi-squared. However, for the Binomial and Poisson distributions, for example, D can be calculated and used directly as a goodness of fit statistic.

Example 3. Deviance for a Binomial model
If the response variables $Y_1, \ldots, Y_N$ are independent and $Y_i \sim Bin(n_i, \pi_i)$.

(i) Find the deviance, D.

(ii) Show that if $Y_i \sim Bin(1, \pi_i)$, D can be reduced to

$$D = -2\left[\sum_{y=1} \ln \hat{y}_i + \sum_{y=0} \ln(1 - \hat{y}_i)\right].$$

Example 4. Drivers are classified as low risk (class 0) and high risk (Class 1). You use a generalized linear model to predict the class. For 6 drivers, the results are

Actual class	0	0	0	1	1	1
Fitted class	0.25	0.35	0.12	0.47	0.84	0.52

Calculate the deviance statistic.

Example 5.
If the response variables $Y_1, \ldots, Y_N$ are independent and $Y_i \sim POI(\lambda_i)$. Find the deviance, D.

Example 6.
A Poisson regression with a log link is run. The results are

Actual	0	1	1	2	2
Fitted	0.35	0.85	1.22	1.74	1.84

Caculate the deviance.

5.7 Hypothesis testing

Hypotheses about a parameter vector $\boldsymbol{\beta}$ of length p can be tested using the sampling distribution of the Wald statistic $(\mathbf{b}-\boldsymbol{\beta})^T\mathcal{J}(\mathbf{b}-\boldsymbol{\beta}) \sim \chi^2(p)$. Occasionally the score statistic is used: $\mathbf{U}^T\mathcal{J}^{-1}\mathbf{U} \sim \chi^2(p)$.

An alternative approach is to compare the goodness of fit of two models. The models need to be nested or hierarchical, that is, they have the same probability distribution and the same link function, but the linear component of the simpler model M_0 is a special case of the linear component of the more general model M_1.
Consider the null hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$$

corresponding to model M_0 and a more general

hypothesis

$$H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_2 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

corresponding to M_1, with $q < p < N$.

We can test H_0 against H_1 using the difference of the deviance statistics

$$\begin{aligned}\Delta D &= D_0 - D_1 \\ &= 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b}_0;\mathbf{y})] \\ &\quad -2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b}_1;\mathbf{y})] \\ &= 2[l(\mathbf{b}_1;\mathbf{y}) - l(\mathbf{b}_0;\mathbf{y})]\end{aligned}$$

If both models describe the data well, then $D_0 \sim \chi^2(N-q)$ and $D_1 \sim \chi^2(N-p)$ so that $\Delta D \sim \chi^2(p-q)$, provided that certain independence conditions hold. So we reject H_0 if

$$\Delta D > \chi^2_\alpha(p-q)$$

or

$$\text{p-value} = P(\chi^2(p-q) > \Delta D) > \alpha.$$

Provided that the deviance can be calculated from the data, ΔD provides a good method for hypothesis testing. The sampling distribution of ΔD is usually better approximated by the chi-squared distribution than is the sampling distribution of a single deviance.

For models based on the Normal distribution, or other distributions with nuisance parameters that are not estimated, the deviance may not be fully determined from the data.

Example 7. The regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ is being investigated. The following maximized log-likelihoods are obtained:

- Using only intercept term:-1126.91
- Using only intercept term, X_1 and X_2: -1122.41
- Using all four terms: -1121.91

The null hypothesis $\beta_1 = \beta_2 = \beta_2 = 0$ is being tested using the likelihood ratio test. Determine the smallest significance level at which you reject the null hypothesis.